Text Analytics: the convergence of Big Data and Artificial Intelligence

Antonio Moreno[1], Teófilo Redondo[2]

[1]*Universidad Autónoma de Madrid and Instituto de Ingeniería del Conocimiento, Madrid, Spain*
[2]*ZED Worldwide, Madrid, Spain*

***Abstract* — The analysis of the text content in emails, blogs, tweets, forums and other forms of textual communication constitutes what we call text analytics. Text analytics is applicable to most industries: it can help analyze millions of emails; you can analyze customers' comments and questions in forums; you can perform sentiment analysis using text analytics by measuring positive or negative perceptions of a company, brand, or product. Text Analytics has also been called text mining, and is a subcategory of the Natural Language Processing (NLP) field, which is one of the founding branches of Artificial Intelligence, back in the 1950s, when an interest in understanding text originally developed. Currently Text Analytics is often considered as the next step in Big Data analysis. Text Analytics has a number of subdivisions: Information Extraction, Named Entity Recognition, Semantic Web annotated domain's representation, and many more. Several techniques are currently used and some of them have gained a lot of attention, such as Machine Learning, to show a semisupervised enhancement of systems, but they also present a number of limitations which make them not always the only or the best choice. We conclude with current and near future applications of Text Analytics.**

***Keywords* — Big Data Analysis, Information Extraction, Text Analytics**

I. Introduction

Natural Language Processing (NLP) is the practical field of Computational Linguistics, although some authors use the terms almost interchangeably. Sometimes NLP has been considered a subdiscipline of Artificial Intelligence, and more recently it sits at the core of Cognitive Computing, since most cognitive processes are either understood or generated as natural language utterances.

NLP is a very broad topic, and includes a huge amount of subdivisions: Natural Language Understanding, Natural Language Generation, Knowledge Base building, Dialogue Management Systems (and Intelligent Tutor Systems in academic learning systems), Speech Processing, Data Mining – Text Mining – Text Analytics, and so on. We will focus here in this specific article in Text Analytics (TA).

Text Analytics is the most recent name given to Natural Language Understanding, Data and Text Mining. In the last few years a new name has gained popularity, Big Data, to refer mainly to unstructured text (or other information sources), more often in the commercial rather than the academic area, probably because unstructured free text accounts for 80% in a business context, including tweets, blogs, wikis and surveys [1]. In fact there is a lack of academic papers covering this topic, although this may be changing in the near future.

Text Analytics has become an important research area. Text Analytics is the discovery of new, previously unknown information, by automatically extracting information from different written resources.

II. Text Analytics: Concepts and Techniques

Text Analytics is an extension of data mining, that tries to find textual patterns from large non-structured sources, as opposed to data stored in relational databases. Text Analytics, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting non-trivial information and knowledge from unstructured text. Text Analytics is similar to data mining, except that data mining tools are designed to handle structured data from databases, either stored as such or as a result from preprocessing unstructured data. Text Analytics can cover unstructured or semi-structured data sets such as emails, full-text documents and HTML files, blogs, newspaper articles, academic papers, etc. Text Analytics is an interdisciplinary field which draws on information extraction, data mining, machine learning, statistics and computational linguistics.

Text Analytics is gaining prominence in many industries, from marketing to finance, because the process of extracting and analysing large quantities of text can help decision-makers to understand market dynamics, predict outcomes and trends, detect fraud and manage risk.

The multidisciplinary nature of Text Analytics is key to understand the complex integration of different expertise: computer engineers, linguists, experts in Law, BioMedicine or Finance, data scientists, psychologists, causing that the research and development approach is fragmented due to different traditions, methodologies and interests.

A typical text analytics application consists of the following steps and tasks:

Starting with a collection of documents, a text mining tool retrieves a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. The underlying strategy in all the components is to find a pattern (from either a list or a previous process) which matches a rule, and then to apply the rule which annotates the text. Each component performs a particular process on the text, such as: *sentence segmentation* (dividing text into sentences); *tokenization* (words identified by spaces between them); *part-of-speech tagging* (noun, verb, adjective, etc., determined by look-up and relationships among words); *shallow syntactic parsing/ chunking* (dividing the text by noun phrase, verb phrase, subordinate clause, etc.); *named entity recognition (NER)* (the entities in the text such as organizations, people, and places); *dependency analysis* (subordinate clauses, pronominal anaphora [i.e., identifying what a pronoun refers to], etc.).

The resulting process provides "structured" or semi-structured information to be further used (e.g. Knowledge Base building, Ontology enrichment, Machine Learning algorithm validation, Query Indexes for Question & Answer systems).

Some of the techniques that have been developed and can be used in the text mining process are information extraction, topic tracking, summarization, categorization, clustering, concept linkage, information visualization, question answering, and deep learning.

DOI: 10.9781/ijimai.2016.369

A. Information Extraction

Information extraction (IE) software identifies key phrases and relationships within text. It does this by looking for predefined sequences in text, a process usually called pattern matching, typically based on regular expressions. The most popular form of IE is named entity recognition (NER). NER seeks to locate and classify atomic elements in text into predefined categories (usually matching preestablished ontologies). NER techniques extract features such as the names of persons, organizations, locations, temporal or spatial expressions, quantities, monetary values, stock values, percentages, gene or protein names, etc. These are several tools relevant for this task: Apache OpenNLP [2], Stanford Named Entity Recognizer [3] [4], LingPipe [5].

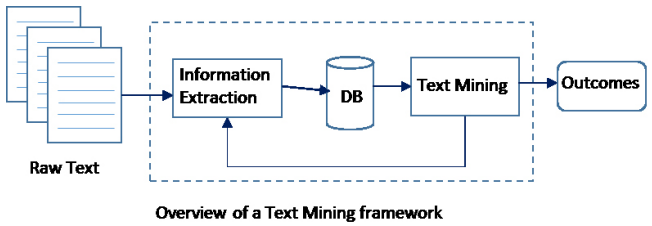

Fig. 1. Overview of a Text Mining Framework

B. Topic Tracking and Detection)

Keywords are a set of significant words in an article that gives a high-level description of its contents to readers. Identifying keywords from a large amount of online news data is very useful in that it can produce a short summary of news articles. As online text documents rapidly increase in size with the growth of WWW, keyword extraction [6] has become the basis of several text mining applications such as search engines, text categorization, summarization, and topic detection. Manual keyword extraction is an extremely difficult and time consuming task; in fact, it is almost impossible to extract keywords manually in case of news articles published in a single day due to their volume.

A topic tracking system works by keeping user profiles and, based on the documents the user views, predicts other documents of interest to the user. Google offers a free topic tracking tool [7] that allows users to choose keywords and notifies them when news relating to those topics becomes available. NER techniques are also used in enhancing topic tracking and detection by matching names, locations or usual terms in a given topic by representing similarities with other documents of similar content [8]. Topic detection is closely related with Classification (see below).

C. Summarization

Text summarization has a long and fruitful tradition in the field of Text Analytics. In a sense text summarization falls also under the category of Natural Language Generation. It helps in figuring out whether or not a lengthy document meets the user's needs and is worth reading for further information. With large texts, text summarization processes and summarizes the document in the time it would take the user to read the first paragraph. The key to summarization is to reduce the length and detail of a document while retaining its main points and overall meaning.

One of the strategies most widely used by text summarization tools is sentence extraction. Important sentences from an article are statistically weighted and ranked. Summarization tools may also search for headings and other markers of subtopics in order to identify the key points of a document.

The methods of summarization can be classified in two broad groups:

- shallow analysis, restricted to the syntactic level of representation and try to extract significant parts of the text;
- deeper analysis, assumes a semantics level of representation of the original text (typically using Information Retrieval techniques).

A relatively recent European Union project, ATLAS, has performed an extensive evaluation of text summarization tools [9].

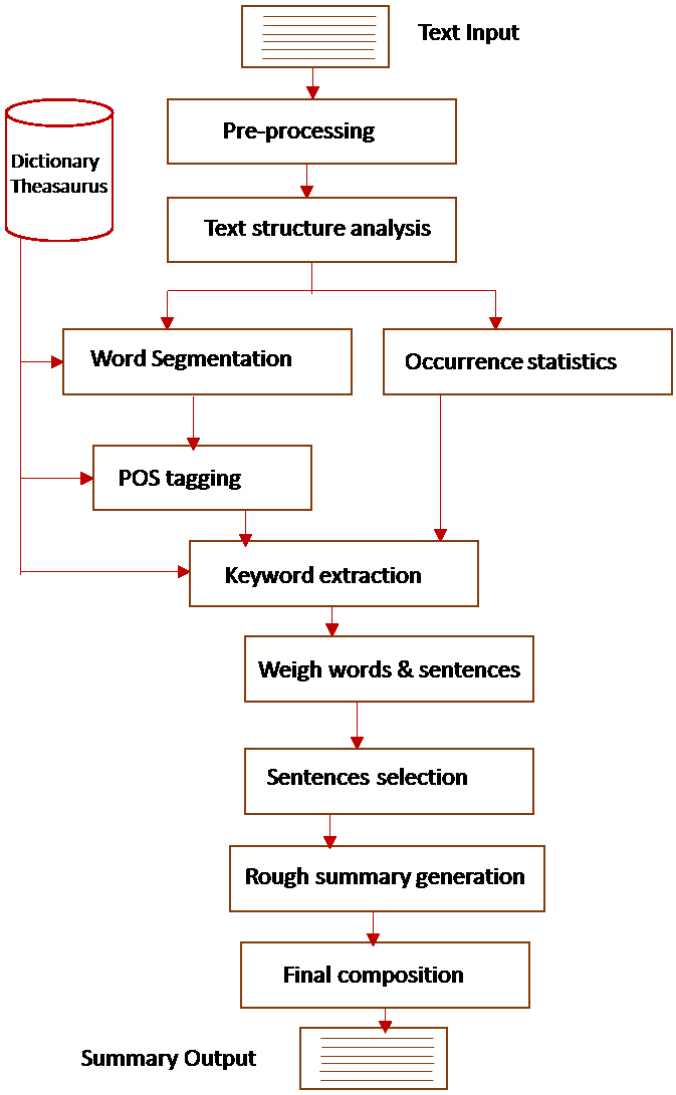

Fig. 2. Text Summarization

A. Categorization or Classification

Categorization involves identifying the main themes of a document by placing the document into a predefined set of topics (either as taxonomies or ontologies). Categorization only counts words that appear and, from the counts, identifies the main topics that the document covers. Categorization often relies on relationships identified by looking for broad terms, narrower terms, synonyms, and related terms. Categorization tools normally have a method for ranking the documents in order of which documents have the most content on a particular topic [10]. Another method is to represent topics as thematic graphs, and using a degree of similarity (or distance from the "reference" graph) to classify documents under a given category [11].

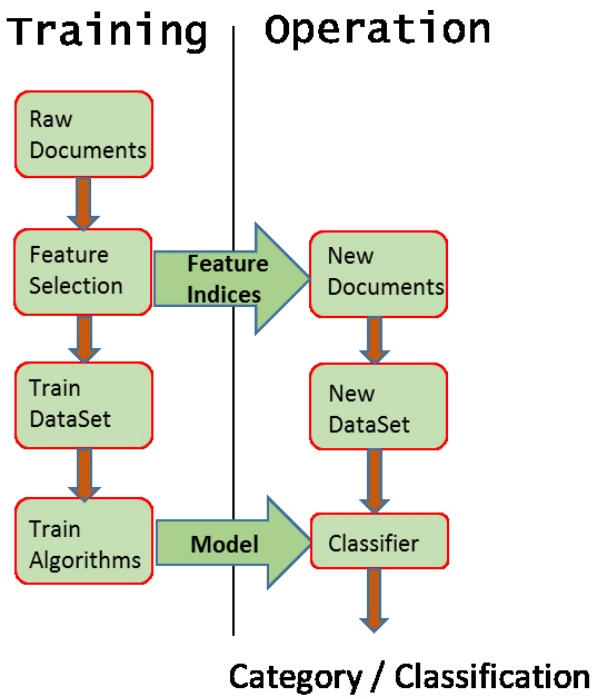

Fig. 3. Text Classification

D. Clustering

Clustering is a technique used to group similar documents, but it differs from categorization in that documents are clustered without the use of predefined topics. In other words, while categorization implies supervised (machine) learning in the sense that previous knowledge is used to assign a given document to a given category, clustering is unsupervised learning: there are no previously defined topics or categories. Using clustering, documents can appear in multiple subtopics, thus ensuring that a useful document will not be omitted from search results (multiple indexing references). A basic clustering algorithm creates a vector of topics for each document and assigns the document to a given topic cluster.

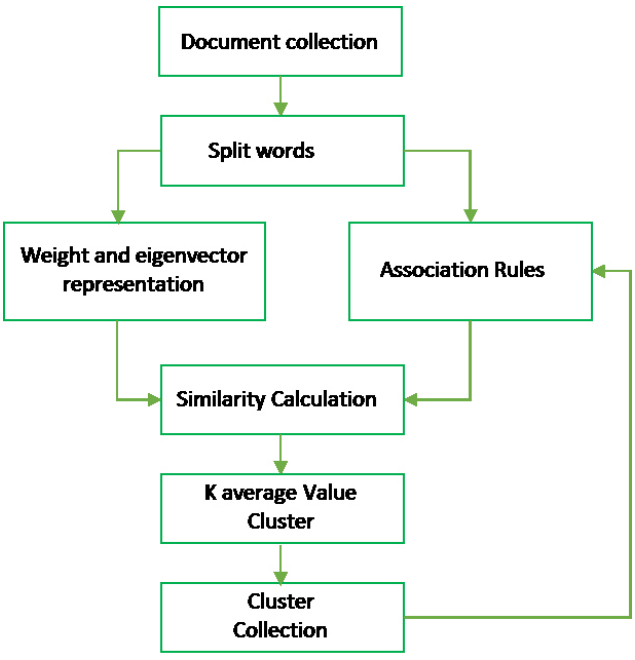

Fig. 4. Document Clustering

Medicine and Legal research papers have been a fertile ground to apply text clustering techniques [12] [13].

E. Concept Linkage

Concept linkage tools connect related documents by identifying their commonly-shared concepts and help users find information that they perhaps would not have found using traditional searching methods. It promotes browsing for information rather than searching for it. Concept linkage is a valuable concept in text mining, especially in the biomedical and legal fields where so much research has been done that it is impossible for researchers to read all the material and make associations to other research.

The best known concept linkage tool is C-Link [14] [15]. C-Link is a search tool for finding related and possibly unknown concepts that lie on a path between two known concepts. The tool searches semi-structured information in knowledge repositories based on finding previously unknown concepts that lie between other concepts.

F. Information Visualization

Visual text mining, or information visualization, puts large textual sources in a visual hierarchy or map and provides browsing capabilities, in addition to simple searching. Information visualization is useful when a user needs to narrow down a broad range of documents and explore related topics. A common typical example of text information visualization are Tag clouds [16], like those provided by tools such as Wordle [17]. Hearst [18] has written an extensive overview of current (and recent past) tools for text mining visualization, but definitively needs an update with the appearance of new tools in recent years: D3.js [19], Gephi [20], as well as assorted JavaScript-based libraries (Raphäel, jQuery Visualize).

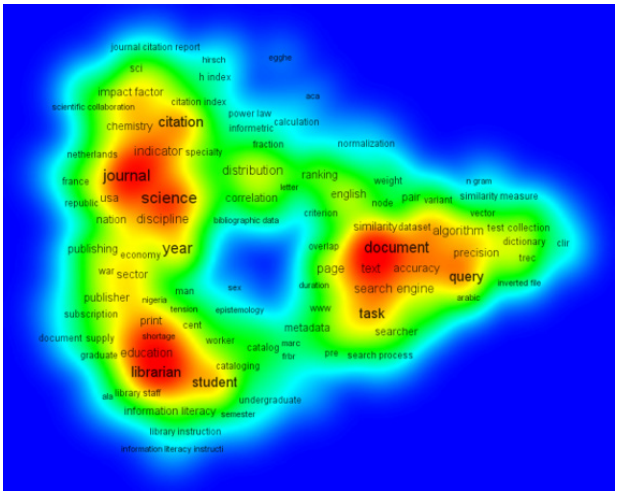

Fig. 5. Text data visualization
(Source: Hearst (2009) “Information Visualization for Text Analysis”)

G. Question Answering

Question answering (Q&A) systems used natural language queries to find the best answer to a given question. Question answering involves a lot of techniques described here, from information extraction for the question topic understanding, question typology and categorization, up to the actual selection and generation of the answer [21] (Hirschman & Gaizauskas, 2001). OpenEphyra [22] was an open-source question answering system, originally derived from Ephyra, which was developed by Nico Schlaefer and has participated in the TREC question answering competition [23]. Unfortunately, OpenEphyra has been discontinued and some alternatives have appeared, such as YodaQA [24], which is general purpose QA system, that is, an “open domain”

general factoid question answering [25].

H. Deep Learning

Deep Learning has been gaining a lot of popularity as of the last two years, and has begun to be experimented for some NLP tasks. Deep Learning is a very broad field and most promising work is moving around Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN).

Neural Networks have a long and prestigious history, and interest within the field of Text Analytics has been revived recently. In a traditional neural network all inputs (and outputs) are independent of each other. The idea behind RNNs is to make use of sequential information (as the words in a sentence). In order to predict the next word in a sentence we must know which words came before it. RNNs perform the same task for every element of a sequence, with the output being dependent on the previous computations. RNNs have a "memory" which captures information about what has been calculated so far. With RNN a given language model can be built, which in turn allows to score arbitrary sentences based on how likely they are to occur in the real world, and later that model allows to generate new text.

CNNs are basically just several layers of convolutions over the input layer to compute the output. Each layer applies different filters, typically hundreds or thousands, and combines their results. These can be used for Sentiment Analysis, Spam Detection or Topic Categorization, but they are of little use with PoS Tagging or Entity Extraction unless additional features are included as filters.

DL4J is a tool for textual analysis using deep learning techniques [26]. It builds on Word2vec, a two-layer neural network that processes text, created by Google [27].

III. Known Problems in Text Analytics

In the context of TA, Big Data is simply a massive volume of written language data. But where does the frontier lie between Big Data and Small Data? There has been a culture-changing fact: while merely 15 years ago a text corpus of 150 million words was considered huge, currently no less than 8.000 million word datasets are available. Not only is it a question simply about size, but also about quality and veracity: data from social media are full of noise and distortion. All datasets have these problems but they are more potentially serious for large datasets because the computer is an intermediary and the human expert do not see them directly, as is the case in small datasets. Therefore, data cleansing processes consume significant efforts and often after the cleansing, the availability of information to train systems is not enough to get reliable predictions, as happened in the Google Flu Trends failed experiment [27].

The reason is that most big datasets are not the output of instruments designed to produce valid and reliable data for analysis, and also because data cleansing is about (mostly subjective) decisions on the relevant design features. Another key issue is the access to the data. In most cases, the academic groups have no access to data from companies such as Google, Twitter or Facebook. For instance, Twitter only makes a very small fraction of its tweets available to the public through its APIs. Additionally, the tweets available do not follow a given pattern (they are an "assorted bunch") so it is difficult to arrive at a conclusion concerning their representativeness. As a consequence, the replication of analyses is almost impossible, since the supporting materials and the underlying technology are not publicly available. Boyd and Crawford [29] go further: limited access to Big Data creates new digital divides, the Big Data rich and the Big Data poor. One needs the means to collect them, and the expertise to analyze them. Interestingly, small but well curated collections of language data (the traditional corpora) offer information that cannot be inferred from big datasets [30].

How to grasp the figurative uses of language, basically irony and metaphor, is also a well-known problem to properly understand text. Essentially, the user's intentions are hidden because the surface meaning is different to the underlying meaning. As a consequence, the words must be interpreted in context and with extra-linguistic knowledge, a fact that being hard on humans, it is even harder for machines. How to translate a given metaphor into another language is extremely difficult. Some estimates calculate that figurative language is about 15-20% of the total content in social media conversations.

IV. Some Use Cases

Text Analytics has produced useful applications for everyday use. Here we just show a sample of these.

A. Lynguo

Social Media Analytics (SMA) consists of gathering data from the social media (Twitter, Facebook, blogs, RSS, websites) dispersed across a myriad of on-line, dynamic, sources; after analyzing automatically the data, these are shown to the client in a graphical environment (dashboard) to help adopting business decisions. Two are the commercial applications: the first one is focused on users' profiles and their network; the other one is targetting the content of the messages. In both cases, the SMA tools support marketing and customer service activities.

Customer profiling is the task of analysing the presence of a brand or the spread of a user's posted content in a social network and extracting information from the users related to gender, age, education, interests, consumer habits, or personality. Typically, the tool provides metrics on influencers and most commented posts. Also, it can identify similarities among users/customers, conversation groups and opinion leaders (Social Intelligence).

Content analytics is the task of analyzing the social media written messages to detect and understand the people's opinions, sentiments, intentions, emotions about a given topic, brand, person. After analyzing millions of comments in almost real-time, these applications help to detect crisis, to measure popularity, reputation and trends, to know the impact of marketing campaigns and customer engagement, or to find out new business opportunities.

Lynguo is an intelligent system developed by the Instituto de Ingeniería del Conocimiento (IIC – http://www.iic.uam.es) that combines both services (the network and the semantics) in a single suite. For instance, Lynguo Observer provides complete metrics on users, comments, trending topics, and their time evolution as well as geolocalization when available. Lynguo Opinion focuses on comments' content: the polarity (positive, neutral or negative) and their distribution through time, and classified by topics and by brands. Lynguo Ideas complements the semantic analysis identifying keywords and concepts associated to brands and people in the social media. This functionality helps to know the distinctive words for a given entity with respect to their competitors based on their exclusive presence or more frequent occurrence of those words in the messages for that entity. Lynguo Alert allows to set up conditions for alerts, for instance, high impact posts or sent by specific users. Finally, Lynguo Report generates personalized reports with graphics from information previously analyzed by other Lynguo modules.

In the next future, two new functionalities will be added: Lynguo Intention and Lynguo Emotions. Analyzing intentions in utterances is an old topic in NLP and AI since the 70s. The goal is to detect what the speaker plans to pursue with his/her speech acts. In the current context of Opinion Mining and Sentiment Analysis [31], intention detection is

focused on determining whether the social media users are expressing a complaint, a question, a wish, a suggestion about a product or a service. The practical application of this knowledge can be useful to the customer services and on-line complaint facilities. Emotions are basically positive or negative, being positive words more frequent, thus carrying less information than negative words [32]. In Lynguo, there are as many as twenty different categories for emotions, from happiness to anger, including intermediate states. The idea is to have a full range of granularity for emotions, and use a different scale for a given task.

There are two main approaches to extraction of opinion polarity in utterances [33]: the statistical or machine learning approach and the lexicon-based approach. The first is the most extended, and basically is a supervised classification task, where the features are learned from annotated instances of texts or sentences. The lexicon-based method involves using a dictionary of words and phrases with polarity values. Then, the program analyzes all relevant polarity words in a given text annotating their scores, and calculates the final aggregation into a final score. The statistical classifiers reach quite a high accuracy in detecting polarity of a text in the domain that they are trained on, but their performance drops drastically when the domain is different. The lexicon-based classifiers operate in a deeper level of analysis, including grammatical and syntactical information for dealing with negation and intensification, for which syntactic parsers are introduced to analyze sentences. Needless to say that Lynguo makes use of this approach.

B. IBM's Watson

IBM has a long history in AI research, and they consider themselves (http://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=1569) as one of the founding fathers of AI in the early days in the 1950s. Along the years IBM created a checkers player [34]; a robotic system trained to assemble some parts of an IBM typewriter [35]; Deep Blue, the specialized chess-playing server that beat then World Chess Champion, Garry Kasparov [36]; TD-Gammon, a backgammon playing program using reinforcement learning (RL) with multiple applications [37]; and pioneering work in Neural Network Learning, inspired in biological information processing [38]. More recently work focused on advanced Q&A (question and answer) and Cognitive Computing, of which the SyNAPSE project and Watson are the more well-known examples.

IBM Watson is a technology platform that uses natural language processing and machine learning to reveal insights from large amounts of unstructured data. Watson can extract information from all kinds of text-rich documents in a repository, it can then build patterns of the relationships of the different concepts in the documents, and then can create a Q&A system to query the contents of the repository. Having been trained using supervised learning techniques, Watson uses natural language processing to understand grammar and context by annotating components to extract mentions and relationships, evaluates all possible meanings and determines what is being asked, and presents answers based on the supporting evidence and quality of information provided.

C. IPsoft's Amelia

IPsoft has been one leading proponent of automatization as part of the future of IT, where more and more menial tasks will be performed by expert systems rather than by people. IPsoft launched an initial version of Amelia, as the software platform is known, in September 2014. A new version, Amelia 2.0, has been presented in the last part of 2015, representing a new generation of this artificial intelligence (AI) platform with the promise that near human cognitive capabilities are ever closer [39].

The platform understands the semantics of language and can learn to solve business process queries like a human being would do. It can also solve technical problems by initially learning the same manuals as humans and then learning through experience and by observing the interactions between human agents and customers using semi-supervised machine learning techniques. Amelia 2.0 can complete more tasks and absorb more knowledge now as its core understanding capabilities mature. The latest version has improvements in dialogue management comprehension and emotional engagement [40]:

- *Memory* – Amelia's declarative memory consists of episodic memory and semantic memory. Episodic memory gives the basic cognition of various experiences and events in time in a sequenced autobiographical form. Semantic memory gives a structured record of facts, meanings, concepts and knowledge about the world/domain.
- *Contextual comprehension* – Concepts and ideas in the human brain are semantically linked. Amelia can quickly and reliably retrieve information across a wider and more complex set of knowledge.
- *Emotional responsiveness* – Research shows that a better user experience is directly tied to empathy shown by the agent throughout the interaction. In addition to an increased emotional quotient (EQ), Amelia presents a mood and a personality vector in a 3-dimensional emotional space.

The software is already used for services such as technology helpdesks, contact centres, procurement processing and to advise field engineers, among other business processes.

Both Watson and Amelia were already briefly mentioned in Redondo [41], as the purpose of both systems is to extend a human's capabilities by applying intelligent artificial systems techniques, such as deep learning and interpersonal communication through dialogue management.

V. Examples of TA Applications

We will briefly review two prominent areas of application of Text Analytics, with a large commercial impact: (1) Medical Analytics – classification of articles of medical content, and (2) Legal Analytics – Information extraction from legal texts.

A. Medical Analytics – Classification of articles or medical content

Biomedical text mining or BioNLP presents some unique data types. Their typical texts are abstracts of scientific papers, as well as medical reports. The main task is to classify papers by many different categories, in order to feed a database (like MEDLINE). Other applications include indexing documents by concepts, usually based or related to ontologies (like Unified Medical Language System-UMLS, or SNOMED-CT) or performing "translational research," that is, using basic biological research to inform clinical practice (for instance, automatically extraction of drug-drug interactions, or gene associations with diseases, or mutations in proteins).

The NLP techniques include biomedical entities recognition, pattern recognition, and machine learning for extracting semantic relations between concepts. Biomedical entities recognition consists of recognizing and categorizing entity names in biomedical domains, such as proteins, genes, diseases, drugs, organs and medical specialties. A variety of lexical resources are available in English and other languages (ontologies, terminological databases, nomenclatures), as well as a wide collection of annotated corpora (as GENIA) with semantic and conceptual relations between entities. Despite their availability, no single resource is enough nor comprehensive since new drugs and genes are discovered constantly. This is the main challenge for BioNLP.

There are three approaches for extracting relations between entities:

- Linguistic-based approaches: the idea is to employ parsers to grasp syntactic structures and map them into semantic representations. They are typically based on lexical resources and their main drawbacks are the abundance of synonyms and spelling variations for entities and concepts.
- Pattern-based approaches: these methods make use of a set of patterns for potential relationships, defined by domain experts.
- Machine Learning-based approaches: from annotated texts by human experts, these techniques extract relations in new collections of similar texts. Their main shortcoming is the requirement of computationally expensive training and testing on large amounts of human-tagged data. To extend the extraction system to another type of data or language requires new human effort in annotation.

Friedman et al. [42] presents a survey of the state of the art and prospects in BioNLP, sponsored by the US National Library of Medicine. This report identifies that "the most significant confounding element for clinical NLP is inaccessibility of large scale de-identified clinical corpora, which are needed for training and evaluation."

B. Legal Analytics – Information extraction from legal texts

One area getting a lot of attention about the practicalities of Text Analytics is that concerning the information extraction from texts with legal content. More specifically, litigation data is full of references to judges, lawyers, parties (companies, public organizations, and so on), and patents, gathered from several millions of pages containing all kinds of Intellectual Property (IP) litigation information. This has given rise to the term Legal Analytics, since analytics helps in discovering patterns with meaning hidden in the repositories of data. What it means to lawyers is the combination of insights coming from bottom-up data with top-down authority and experience found in statutes, regulations and court sentences. All this places objective data at the center instead of the so-called anecdotal data.

The underlying problem is that legal textual information is expressed in natural language. While a search can be made for the string *plaintiff*, there are no searches for a string that represents an individual who bears the role of plaintiff. To make language on the Web more meaningful and structured, additional content must be added to the source material, which is where the Semantic Web (semantic roles' tagging) and Natural Language Processing perform their contribution.

We start with an input, the corpus of texts, and then an output, texts annotated with XML tags, JSON tags or other mechanisms. However, getting from a corpus of textual information to annotated output is a demanding task, generically referred to as the knowledge acquisition bottleneck [43]. This task is very demanding on resources (especially manpower with enough expertise to train the systems) and it is also highly knowledge-intensive since whoever is doing the annotation must know what and how to annotate knowledge related to a given domain.

Processing Natural Language (NL) to support such richly annotated documents presents some inherent issues. NL supports all of the following, among other things:

(1) *implicit or presupposed information* – "When did you stop taking drugs?" (presupposes that the person is questioned about taking drugs at some time in the past);

(2) *multiple forms with the same meaning* – Jane Smith, Jane R. Smith, Smith, Attorney Smith… (our NER system must know that these are different ways to refer to the same physical person);

(3) *the same form with different contextually dependent meanings* – An individual referred to as "Jane Smith" in one case decision may not be the individual referred to by the name "Jane Smith" in another case decision; and

(4) *dispersed meanings* – Jane Smith represented Jones Inc. She works for Boies, Schiller and Flexner. To contact her, write to j.smith@bsfllp.com

People grasp naturally relationships between words and phrases, such that if it is true that *Bill used a knife to injure Phil*, then *Bill injured Phil*. Natural Language Processing (NLP) addresses this highly complex problem as an *engineering* problem, decomposing large problems into smaller problems and subdomains (e.g. summarization, information extraction, …) that we have already covered above. In the area of Legal Analytics we are primarily interested in information extraction. Typically a legal analytics system will annotate elements of interest, in order to identify a range of particular pieces of information that would be relevant to legal professionals such as:

- Case citation
- Names of parties
- Roles of parties, meaning *plaintiff* or *defendant*
- Type of court
- Names of judges
- Names of attorneys
- Roles of attorneys, meaning the side they represent (plaintiff or defendant)
- Final decision
- Cases cited
- Nature of the case, meaning using keywords to classify the case in terms of subject (e.g., criminal assault, intellectual property, etc.)

The business implications of Legal Analytics have originated a full branch of textual Big Data applications. Some companies have benefitted from a lucrative market, such as LexisNexis (http://www.lexisnexis.com/en-us/gateway.page), focused on offering predictions on potential medical-malpractice cases to specialized attorneys. Quite recently, LexisNexis has acquired Lex Machina [44], a company that mines mainly litigation data around IP information. Lex Machina originated in the departments of Law and Computer Science of Stanford University. Their Legal Analytics Platform helps to create a well-documented strategy to win cases on Patent, Trademark and Copyright data.

Every day, Lex Machina's crawler extracts data (and indexes documents) from several U.S Law repositories. The crawler automatically captures every docket event and downloads key case documents. It converts the documents by optical character recognition (OCR) to searchable text and stores each one as a PDF file. When the crawler finds a mention of a patent, it fetches information about that patent from the Patents and Trademarks Office (PTO) site. The crawler invokes Lexpressions, a proprietary legal text classification engine. The NLP technology classifies cases and dockets and resolves entity names (using a NER engine). A process of curation of the information extracted is performed by specialized attorneys to ensure high-quality data. The structured text indexer then performs a data cleansing operation to order all the data and stores it for search. Lex Machina's web-based application enables users to run search queries that deliver easy information retrieval of the relevant docket entries and documents.

Most if not all of the work around either medical or legal text analytics has originated from data sets in English (the above-mentioned MEDLINE, LexisNexis, or others like MedScape), with little to no work in other languages. This could be an opportunity to create new text analytics systems in languages other than English. For instance, considering the fact that the legal systems vary quite significantly from

country to country, this is a good field to grow new areas of business in the so-called Digital Economy (see also IJIMAI's last year special issue on this topic).

VI. Future Work

The technologies around text analytics are currently being applied in several industries, for instance, sentiment and opinion analysis in media, finance, healthcare, marketing branding or consumer markets. Insights are extracted not only from the traditional enterprise data sources, but also from online and social media, since more and more the general public has turned out to be the largest generator of text content (just imagine online messaging systems like Whatsapp or Telegram).

The current state of text analytics is very healthy, but there is room for growth in areas such as customer experience, or social listening. This bears good promises for both scientific experimentation and technical innovation alike: Multi-lingual analytics is facilitated by machine learning (ML) and advances in machine translation; customer experience, market research, and consumer insights, and digital analytics and media measurement are enhanced through text analytics; besides the future of deep learning in NLP, long-established language-engineering approaches taxonomies, parsers, lexical and semantic networks, and syntactic-rule systems will continue as bedrocks in the area; emotion analytics, affective states compounded of speech and text as well as images and facial-expression analysis; new forms of supratextual communications like emojis need their own approach to extract semantics and arrive at meaningful analytics; semantic search and knowledge graphs, speech analytics and simultaneous machine translation; and machine-written content, or the capability to compose articles (and email, text messages, summaries, and translations) from text, data, rules, and context, as captured previously in the analytics phase.

VII. Conclusion

Text Analytics, with its long and prestigious history, is an area in constant evolution. It sits at the center of Big Data's Variety vector, that of unstructured information, especially with social communications, where content is generated by millions of users, content not only consisting of images but most of the times textual comments or full blown articles. Information expressed by means of texts involves lots of knowledge about the world and about the entities in this world as well as the interactions among them. That knowledge about the world has already been put to use in order to create the cognitive applications, like IBM's Watson and IPsoft's Amelia, that will interact with human beings expanding their capabilities and helping them perform better. With increased communication, Text Analytics will be expanded and it will be needed to sort out the noise and the irrelevant from the really important information. The future looks more than promising.

References

[1] Xerox Corporation (2015): http://www.xrce.xerox.com/Research-Development/Industry-Expertise/Finance (accessed 26 December 2015)

[2] Apache OpenNLP (2015): http://opennlp.apache.org/ (accessed 19 December 2015)

[3] Stanford Named Entity Recognizer (2015): http://www-nlp.stanford.edu/software/CRF-NER.shtml (accessed 19 December 2015)

[4] J. R. Finkel, T. Grenager, and C. Manning (2005). "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling". *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*. (online reading: http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf)

[5] LingPipe (2011): http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html (accessed 29 November 2015)

[6] S. Lee and H. Kim (2008). "News Keyword Extraction for Topic Tracking". *Fourth International Conference on Networked Computing and Advanced Information Management*, IEEE.

[7] Google Alerts (2016): http://www.google.com/alerts (accessed 10 January 2016)

[8] W. Xiaowei, J. Longbin, M. Jialin and Jiangyan (2008). "Use of NER Information for Improved Topic Tracking", *Eighth International Conference on Intelligent Systems Design and Applications*, IEEE Computer Society.

[9] ATLAS Project (2013): http://www.atlasproject.eu/atlas/project/task/5.1 (accessed 10 January 2016)

[10] G. Wen, G. Chen, and L. Jiang (2006). "Performing Text Categorization on Manifold". *2006 IEEE International Conference on Systems, Man, and Cybernetics*, Taipei, Taiwan, IEEE.

[11] H. Cordobés, A. Fernández Anta, L.F. Chiroque, F. Pérez García, T. Redondo, A. Santos (2014). "Graph-based Techniques for Topic Classification of Tweets in Spanish". *International Journal of Interactive Multimedia an Artificial Intelligence*.

[12] T. Theodosiou, N. Darzentas, L. Angelis, C.A. Ouzonis (2008). "PuReD-MCL: a graph-based PubMed document clustering methodology". *Bioinformatic*s 24.

[13] Q. Lu, J. G. Conrad, K. Al-Kofahi, W. Keenan (2011). "Legal document clustering with built-in topic segmentation", *Proceedings of the 20th ACM international conference on Information and knowledge management*.

[14] P. Cowling, S. Remde, P. Hartley, W. Stewart, J. Stock-Brooks, T. Woolley (2010), "C-Link Concept Linkage in Knowledge Repositories". *AAAI Spring Symposium Series*.

[15] C-Link (2015): http://www.conceptlinkage.org/ (accessed 10 December 2015)

[16] Y. Hassan-Montero, and V Herrero-Solana (2006). "Improving Tag-Clouds as Visual Information Retrieval Interfaces", *I International Conference on Multidisciplinary Information Sciences and Technologies*, InSciT2006.

[17] Wordle (2014): http://www.wordle.net/ (accessed 20 December 2015)

[18] M. A. Hearst (2009) "Information Visualization for Text Analysis", in *Search User Interfaces*. Cambridge University Press (online reading: http://searchuserinterfaces.com/book/)

[19] D3.js (2016): http://d3js.org/ (accessed 20 January 2016)

[20] Gephi (2016) https://gephi.org/ (accessed 20 January 2016)

[21] L. Hirschman, R. Gaizauskas (2001), "Natural language question answering: the view from here", *Natural Language Engineering 7*. Cambridge University Press. (online reading: http://www.loria.fr/~gardent/applicationsTAL/papers/jnle-qa.pdf)

[22] OpenEphyra (2011): https://mu.lti.cs.cmu.edu/trac/Ephyra/wiki/OpenEphyra (accessed 5 January 2016)

[23] N. Schlaefer, P. Gieselmann, and G. Sautter (2006). "The Ephyra QA system". *2006 Text Retrieval Conference (TREC)*.

[24] YodaQA (2015): http://ailao.eu/yodaqa/ (accessed 5 January 2016)

[25] P. Baudis (2015) "YodaQA: A Modular Question Answering System Pipeline". *POSTER 2015 — 19th International Student Conference on Electrical Engineering*. (online reading: http://ailao.eu/yodaqa/yodaqa-poster2015.pdf)

[26] DL4J (2015): http://deeplearning4j.org/textanalysis.html (accessed 16 December 2015)

[27] Google – Word2vec (2013): http://arxiv.org/pdf/1301.3781.pdf (accessed 20 December 2015)

[28] D. Lazer, R. Kennedy, G. King, and A. Vespignani (2014). "Big data. The parable of Google Flu: traps in big data analysis." *Science*, 343(6176).

[29] D. Boyd, and K. Crawford (2011). "Six Provocations for Big Data". *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*. (Available at SSRN: http://ssrn.com/abstract=1926431 or http://dx.doi.org/10.2139/ssrn.1926431)

[30] A. Moreno, and E. Moro (2015). "Big data versus small data: the case of 'gripe' (flu) in Spanish". *Procedia, Social and Behavioral Sciences*, 198.

[31] B. Liu (2012). *Sentiment Analysis and Opinion Mining*. Morgan and Claypool. Chicago.

[32] D. Garcia, A. Garas, and F. Schweitzer (2012). "Positive words carry less information than negative words". *EPJ Data Science*, 1:3. (online reading: http://www.epjdatascience.com/content/1/1/3)

[33] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede (2011).

"Lexicon-based Methods for Sentiment Analysis". *Computational Linguistics*, 37, 2. (online reading: https://www.aclweb.org/anthology/J/J11/J11-2001.pdf)

[34] A. Samuel (1959). "Some Studies in Machine Learning Using the Game of Checkers", *IBM Journal* **3** (3).

[35] C. A. Pratt (1982): "Robotics at IBM", *SIMULATION*.

[36] M. Campbell, A. Hoane, and F. Hsu (2001). "Deep Blue". *Artificial Intelligence,* 134.

[37] G. Tesauro (1995). "Temporal difference learning and TD-Gammon". *Communications of the ACM*, Vol. 38, No. 3.

[38] R. Linsker (1990). "Perceptual Neural Organization: Some Approaches Based on Network Models and Information Theory", *Annual Review of Neuroscience*, Vol. 13.

[39] Karl Flinders (2015). "Amelia, the IPsoft robot". *Computer Weekly* (http://www.computerweekly.com/news/4500254989/Amelia-the-IPsoft-robot-gets-a-makeover

[40] IPsoft (2015) (http://www.ipsoft.com/ipsoft-humanizes-artificial-intelligence-with-the-next-generation-of-its-cognitive-agent-amelia/

[41] T. Redondo (2015). "The Digital Economy: Social Interaction Technologies – an Overview". *International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 3- 2.*

[42] C. Friedman, T. Rindflesch, and M. Corn (2013). "Natural language processing: State of the art and prospect for significant progress, a workshop sponsored by the National Library of Medicine". *Journal of Biomedical Informatics*, 46.

[43] S. Potter (2002). *A Survey of Knowledge Acquisition from Natural Language*, part of the Advanced Knowledge Technologies project, University of Edinburgh.

[44] Lex Machina (2015): https://lexmachina.com/ (accessed 20 December 2015)

Antonio Moreno-Sandoval (BA 1986, MA 1988, PhD 1991, Universidad Autónoma de Madrid, UAM) is Professor of Linguistics and Director of the Computational Linguistics Lab at UAM. He is a former Fulbright postdoc scholar at the Computer Science Dept., New York University (1991-1992) and a former DAAD scholar at Augsburg Universität (1998). His training in Computational Linguistics began as a research assistant in the Eurotra Machine Translation Project (EU FP-2) and then at IBM Scientific Center in Madrid (1989-1990). He was the principal researcher of the Spanish team in the C-ORAL-ROM Project (EU FP-5). He has managed over 15 projects (national, regional-funded) as well as industry contracts. Since 2010 he is Senior Researcher at the Instituto de Ingeniería del Conocimiento (IIC-UAM) in the Social Business Analytics group. Moreno-Sandoval has supervised 9 theses to completion. He is author or co-author of 4 books and over 80 scientific papers.

Teófilo Redondo (BArts -1985, MArts - 1986; Universidad Complutense de Madrid - UCM) is Project Portfolio Coordinator at Zed Worldwide, in the Deparment of Innovation. He was before Technology Architect & Director of Innovation Projects at Universidad Internacional de La Rioja (UNIR). Previously he developed a career at IBM covering several areas like Cloud Computing and Big Data Architectures, Enterprise Solutions Architect (SAP, Oracle Solutions, Dassault Systemes), and as SOA Architect. He started in the IBM Research Division (IBM Scientific Center in Madrid) with several projects on Machine Translation, during which he produced a number of articles on this subject. He was Visiting Scholar at Stanford University (1987). He currently holds an Associate Professorship at UNIR teaching about eLearning in Social Networks as well as Natural Language Processing techniques. He is affiliated with SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural) since almost the beginning.

.